PPDG – Quarterly Status Report from the Caltech Group
September 2000

Julian Bunn, Mehnaz Hafeez, Takako Hickey, Koen Holtman, Iosif Legrand, Vladimir Litvin, Harvey Newman, Asad Samar,

Introduction

The Caltech group is active in several of the PPDG project areas:

- Data Grid Development
  - High throughput data transfer (JB, HN, AS)
  - Globus Security and Information Infrastructure (AS, MH)
  - Tier2 Center Design (HN, JB)
  - Distributed Data Management (JB, HN, AS, MH, KH)
  - Distributed Computing and Task Scheduling (KH, TH, VL, AS, MH)
  - Data Structures and Clustering (KH, JB, HN)

- Distributed System Simulations (IL; HN, KH)
  - Site and Network Configuration and Throughput Studies
  - Production System and ODBMS Studies and Optimization
  - Tape System requirements and Usage-Modes Study
  - Data Access and Analysis Strategy Studies

These topics are covered in detail in the following sections.

High Throughput Data Transfer

There are ongoing tests of high performance networking between the various sites (Caltech, FNAL, CERN, and several other sites active in CMS software and computing), most notably between SLAC and Caltech in relation to the 100 MB/sec PPDG milestone. It is hoped to use ORCA (Object Reconstruction for CMS Analysis) Objectivity database files (as well as BaBar files) once the network is tuned. Coordination between networking experts at all the collaborating institutes is important to ensure that TCP/IP capabilities match, and routes etc. are set correctly. We have been using the NTON Caltech-SLAC link on which we have a dual OC12 connection between a machine in CACR and two machines at SLAC.

It turns out to be difficult to even get iperf or ttcp rates up to the necessary > 800 Mbits/sec that would encourage one to attempt a real file transfer. We currently see an aggregate of a little more than 600 Mbits/sec on NTON, with some mysterious packet retransmits that we are investigating.

CACR has recently ordered new equipment, including a Juniper M160 Router. This will allow us to move from ATM to Gbit Ethernet, and to route traffic between Caltech, JPL, SLAC and other sites over NTON at OC-48 (2.5 Gbps), with the ability to upgrade our setup at CACR to OC-192 in the future. We will then use dual Gbit Ethernet connections across NTON. At some point we will get > 800 MBits/sec and then the challenge is to:

a) choose our Objectivity database files,
b) decide which direction they need to go in,
c) put them on a disk array that can sustain > 100 MBytes/sec read
d) pump them across NTON to a disk array that can sustain > 100 MBytes/sec write

We were initially enthusiastic to use all eight of the OC3 adapters on the Caltech X class Exemplar, which were multiplexed into the two OC12 lines to SLAC. This turned out to be too slow. Only the OC3 adapter on the primary Exemplar node could achieve a substantial fraction of the nominal 155 Mbit/sec, and the maximum rate achievable from the secondary nodes (where data was routed across the Exemplar's inter-node buses) was unacceptably low. After initial tests by Davide Salomoni and discussion with Alex Szalay (JHU) on the I/O capabilities of various servers, we purchased a Dell PowerEdge 4400 with dual 860MHz CPUs, 1GB of RAM, twin OC12 ATM cards, and very fast disk arrays (capable of at least 150 MBytes/sec), running Windows 2000 Server.

We chose to run Windows 2000 initially because its TCP/IP stack is highly efficient (as our tests with Gbit ethernet on a LAN confirm) as it takes maximum advantage of all the hardware capabilities of latest generation network cards (we have had excellent results with those from SysKonnect).

Globus Security and Information Infrastructure

The Caltech group has been actively involved in many Grid-related activities in Europe. We have been working with the Data Management work-package [1] team (WP2) of the EU DataGrid project [2], in the initial design and requirement specification phase. We evaluated real use-cases of the HEP community at CERN and incorporated these in the functionality that this work package offers. This research will appear in our paper "Data Management in an International DataGrid" [3]. AS has also been actively participating in the CMS-related Grid activities. We carried out a project called "Grid Data Management Pilot" (GDMP) [4] which is supposed to fulfill CMS's urgent needs of a DataGrid infrastructure and at the same time act as a pilot for the longer term EU DataGrid project. The first version of this software has been released (Version 1.1) and its design and architecture will be presented in the coming ACAT 2000 workshop at FermiLab.

[1] http://cern.ch/grid-data-management
[2] http://grid.web.cern.ch/grid/
[3] IEEE, ACM International Workshop on Grid Computing [Grid'2000], 17-20 Dec. 2000, Bangalore, India
[4] http://cmsdoc.cern.ch/cms/grid

Tier2 Center Design

Caltech and UCSD are preparing a plan for Tier2 prototypes and Tier1 interaction, which will involve the purchase and installation of hardware and software. ORCA database file replication between CERN, FNAL and the prototype Tier2 servers at Caltech and UCSD will be one of the first tasks. The database files are each typically several hundred MBytes in size. The Tier2 prototypes will probably offer ~2 TByte of online disk storage. It is hard at this stage to estimate accurately the WAN traffic from CERN or FNAL to the Tier2 servers. However, we can postulate a half-fill of the available capacity at each site over a couple of days at the start of an analysis or re-reconstruction task, i.e. an average of ~50 Mbits/sec to both sites, followed by replication between the two sites to fill the remainder of available capacity. This second phase will soak up available bandwidth on the SDSC-Caltech link.

Use of the Tier2 for CMS simulated event production, distribution and analysis will involve groups at UC Davis, UC Riverside and UCLA, as well as Caltech and UCSD. The "California Tier2" concept of a distributed center linked over CALREN2 and NTON will be further developed during 2001, based on fund sharing at some of the university sites mentioned above.

Distributed Computing and Data Management

Data Structures and Clustering

Objectivity-container and user-collection transport R&D at Caltech is focussing on CMS (and DPOSS astrophysics) data. A prototype replication system using new algorithms and based on ORCA, Globus, and PPDG middleware will be ready in time for a demonstration at SuperComputing 2000 (Dallas) in November. The demonstration will involve transparent data clustering and replication for access and processing with improved throughput between the SC2000 conference site, Caltech, CERN and potentially other sites. The Caltech – Dallas path will be instrumented to support data transfers in the Gbps range (up to OC-48). This R&D is progressing in close collaboration with the Globus team of Ian Foster et al., and Johns Hopkins University team of Alex Szalay et al.

Work is in progress on developing software tools and middle-ware that allows for large scientific datasets to be managed and replicated at the granularity level of object collections, and eventually single objects. Efficient and convenient support for data extraction and replication at the level of individual objects and events will enable types of interactive data analysis that would be too inconvenient or costly to perform with tools that work at the file level only. Initial designs for object level tools were presented in February 2000 at the CHEP 2000 conference [1]. A design for a first prototype was completed in July 2000 [2] [3]. Coding began in August 2000. The initial prototype will replicate CMS ORCA physics objects, will use GLOBUS middleware [4] for security and fast data transport, and Objectivity/DB [5] as the underlying storage layer. First results will be presented at ACAT'2000 [6] in October, and the prototype will be demonstrated at Supercomputing 2000 [SC2000] in November. A visit has been made to the SDSS Science Archive team at JHU [7] to exchange knowledge and experience in implementing large science archives using Objectivity/DB. Development copies of the SX query tool and server [8] have been successfully installed at Caltech in July 2000. This is a first step towards installing a complete replica of the production SDSS SX data at Caltech. Koen Holtman and Asad Samar have had frequent contacts with the Globus team [9], to support the creation of the requirements for Globus DataGrid components [10] like the Globus Replica Catalog and Globus Replica Manager [10].

CMS Production

CMS is undertaking a large Monte Carlo simulation and reconstruction production run in Fall 2000, with of order 2 to 4 million events planned to be generated, simulated, reconstructed and then analyzed by several different physics groups. The processing of each event involves several stages, each to be performed at different locations, primarily Caltech, Wisconsin, FNAL and CERN. The processed events will be accessed and analyzed by physicists in those and several other locations. This task will be supported by the Globus-based ORCA file replication services being developed by researchers at Caltech and CERN described above, and in collaboration with the European commision DataGrid project. These services will be implemented as a first prototype in time for the fall 2000 production. The prototype will allow replication of the data and meta data in streaming or on-demand modes Once replicas of the produced events have been made, additional processing steps will be executed at the primary sites, followed by further replication of the new results. At that stage, results can be analyzed by the distributed groups of CMS physicists.

So far, the main focus has been on ORCA4 and CMSIM installation on the CALTECH and Wisconsin facilities:

- *jasper.cacr.caltech.edu*: A SUN Ultra-250 dual CPUmachine at CACR, running the full CMS ORCA and User Analysis Environment. It has been used as a test-bed for the first versions of the automatic file transfer system, which will be used in the Fall 2000 September CMS production. The GLOBUS toolkit has been installed and tested. This will be used together with the Grid-enabled GDMP application as part of a next-generation system for CMS data production and distribution.
- *X-Class Exemplar* at Caltech/CACR - Full version of the automatic file transfer system has been installed together with an adapted CMSIM 116 (CMS simulation program) version. After the CMSIM 120 release (which includes a complete representation of the latest all-Silicon tracker), we will adapt the program for massive simulation production runs (using 240CPUs) in support of the Higher Level Trigger studies which are a major focus of CMS work this year and 2001. 250,000 minimum bias events using the CMSIM 116 version have been generated and stored on the CALTECH HPSS system by means of this system. An automatic event-transferring system (which is to be replaced by the application based on GDMP; see above) was extensively tested during this production run.
- *Linux part of Condor cluste*r at Wisconsin: A full version of the automatic file transfer system has been installed together with recent CMSIM versions. The ORCA 4.2.0 release was successfully used to build and populate Objectivity/DB federations of simulated hits and digits (ooHits & ooDigis); both in a standalone application and in jobs running on the Condor flock (without check-pointing). Approximately 0.5M QCD background events were produced during these runs. We are planning to use the Linux part of the Condor flock at Wisconsin for the Fall CMS production. We also have plans to utilize the Solaris part of Condor and make changes inside ORCA 4.2.0 (and future releases) to enable it to run smoothly in the Condor flock. Additional disk space has been installed on Condor for this purpose.
- *naegling.cacr.caltech.edu* - CMSIM/ORCA 4.2.0 has been installed on this Beowulf-class cluster. We tested managed queuing systems, which were developed in the GIOD[11] project framework. That was successfully tested together with CMSIM 118, and very soon we will make tests with ORCA 4.2.0. 70GB of additional disk space has been installed for that purpose. Work has begun on using naegling as a test-bed for our future Linux PC Farm. The intent is to verify all three major components together - ORCA itself, the automatic file transfer system and the managed queuing system from the GIOD project.
- *future PC Farm*. We have plans to create a test-bed for distributed computations between Caltech, SDSC and CERN, after purchase of the Tier2 PC Farm in October 2000 (see Section on Tier 2 Centers). An upgraded US-CERN link, with a bandwidth of 155Mbps (planned by December 2000), will be used for this purpose.

Transparent migration of data in and out of the Caltech HPSS system will be added as part of the production procedures in the future, based on work to be done in collaboration with EU DataGrid WP5 (on Mass Storage System integration) starting in November 2000.

[1] K. Holtman, H. Stockinger. Building a Large Location Table to Find Replicas of Physics Objects. Proceedings of CHEP 2000, Padova, Italy. http://kholtman.home.cern.ch/kholtman/olt_long.ps
[2] http://www.cacr.caltech.edu/ppdg/meetings/ppdg_collab/holtman/objrepl.pdf
[3] http://kholtman.home.cern.ch/kholtman/globusretreat_objrepl.ppt
[4] www.globus.org

[5] www.objy.com
[6] http://conferences.fnal.gov/acat2000/ [SC2000] http://sc2000.org/
[7] http://www.globus.org/datagrid/
[8] http://www.globus.org/datagrid/deliverables/default.asp
[9] http://www.sdss.jhu.edu/
[10] http://www.sdss.jhu.edu/ScienceArchive/doc.html Resource/Job Management Services
[11] http://pcbunn.cacr.caltech.edu/

Distributed Task Scheduling

In the past quarter a job management service has been developed and prototyped on the *naegling* Linux cluster described above. The service allows clients to submit, monitor, and terminate jobs as a set. It has a scheduling mechanism (to be refined) that allows selection of processors based on processor type, load, available data sets, etc. The service maintains replicated states, so that computations will not be lost if a server fails. The state tolerates network partition failures, so clients will not lose long running jobs when a partition heals. The original implementation of the service [4] used the group communication toolkit Ensemble [2,3]. As the service software was ported to the 65 processor naegling cluster at CACR [1] some scalability problems were encountered. The system was thus redesigned to use group communication only for a small set of servers. To preserve a consistent membership and replication state an additional mechanism was employed, termed reliable RPC. The new design runs well over 65 servers. The earlier version that ran on up to 32 servers was tested with ORCA production software. A paper describing the new design is currently in preparation [5].

[1] Intel Pentium Pro Beowulf Cluster (naegling), http://www.cacr.caltech.edu/resources/naegling
[2] Kenneth P. Birman, "Building Secure and Reliable Network Applications", Manning Publishing Company and Prentice Hall, Jan 1997.
[3] Mark Hayden, "The Ensemble System", Ph.D. thesis, Cornell University, Jan 1998.
[4] Takako M. Hickey and Robbert van Renesse, "An Execution Service for a Partitionable Low Bandwidth Network", In Proceedings of the Twenty-Ninth International Symposium on Fault-Tolerant Computing, Madison, Wisconsin, USA, June 1999. (Also avialable as http://www.hep.caltech.edu/~takako/pubs/pex_ftcs.ps)
[5] Takako M. Hickey, "Augmenting Group Communication to Handle Membership of Larger Groups", in preparation.

Distributed System Simulations

The Caltech group continued the development of the MONARC [1][2] simulation toolset and its validation by simulating the CMS - High Level Trigger Farm (Spring 2000) setup. This first attempt to simulate a large production farm based on Objectivity to store data provided encouraging results in understanding and optimizing large scale distributed systems [3]. Dedicated modules were developed for the simulation framework to allow the study of cost effective solutions in using tapes for different access patterns [4]. Efficient job scheduling policies in very large distributed systems, which evolve dynamically, as the off-line data processing for LHC experiments, is a challenging task. Currently, we are evaluating a possible approach for such a scheduling middle-layer system as a self organizing Neural Network structure which is based on competitive learning from past experience, and which evolves dynamically while trying to optimize the resource utilization and the efficiency for those jobs of high priority.

[1] http://www.cern.ch/MONARC/sim_tool/
[2] http://www.cern.ch/clegrand/MONARC/WSC/monarc_wsc2000.pdf
http://www.cern.ch/clegrand/MONARC/CHEP2k/sim_chep.pdf
[3] http://www.cern.ch/MONARC/sim_tool/Publish/CMS/publish/
http://www.cern.ch/MONARC/sim_tool/Publish/CMS/publish/sim_cms_hlt.pdf
[4] http://www.cern.ch/MONARC/sim_tool/Publish/TAPE/publish/